The Journal of Open Source Software

{fairmetrics}: An R package for fairness evaluation metrics with confidence intervals

Benjamin Smith[1], **Jianhui Gao**[1], **Benson Chou**[1], **and Jessica Gronsbell**[1]

1 University of Toronto

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

Fairness is a growing area of machine learning (ML) that focuses on ensuring models do not produce systematically biased outcomes for certain groups, particularly those defined by protected attributes such as race, gender, or age. Evaluating fairness is a critical aspect of model development, as biased models can perpetuate or exacerbate existing social inequalities. The {fairmetrics} R package offers a user-friendly framework for rigorously evaluating numerous group-based fairness criteria, including metrics based on independence (e.g., statistical parity), separation (e.g., equalized odds), and sufficiency (e.g., predictive parity). These criteria assess whether a model is equally accurate or well-calibrated across predefined groups so that appropriate bias mitigation strategies can be implemented. {fairmetrics} provides both point and interval estimates for multiple metrics through convenient wrapper functions, and includes example an dataset derived from the Medical Information Mart for Intensive Care, version II (MIMIC-II) database (Goldberger et al. 2000; J. Raffa 2016).

Statement of Need

Machine learning (ML) offers significant potential for predictive modelling in biomedical research (Rajpurkar et al. 2022). Despite its promise, there is substantial evidence that, without appropriate forethought and planning, ML models can introduce or exacerbate health inequities by making less accurate decisions for certain groups or individuals (Grote and Keeling 2022). While existing software can compute fairness metrics, none provide out-of-the-box statistical inference, leaving practitioners without guidance on the uncertainty around those metrics. As ML becomes increasingly embedded in healthcare systems, ensuring equitable model performance across diverse populations is essential(Gao et al. 2024). The {fairmetrics} R package fills this gap by offering a suite of popular group-fairness metrics along with bootstrap-based confidence intervals, enabling more rigorous and interpretable assessments of fairness in biomedical ML.

Fairness Criteria

Group fairness criteria are typically classified into three main categories: independence, separation, and sufficiency (Barocas, Hardt, and Narayanan 2023; Berk et al. 2018; Castelnovo et al. 2022). The {fairmetrics} package computes a range of group fairness metrics together with bootstrap-based confidence intervals for uncertainty quantification. The metrics implemented in the package are briefly described below.

Independence

- **Statistical Parity:** Compares the overall rate of positive predictions between groups, irrespective of the true outcome.
- **Conditional Statistical Parity:** Restricts the comparison of positive prediction rates to a specific subgroup (e.g., within a hospital unit or age bracket), offering a more context-specific fairness assessment.

Separation

- **Equal Opportunity:** Compares disparities in false negative rates between two groups, quantifying any difference in missed positive cases.

- **Predictive Equality:** Compares false positive rates (FPR) between groups, ensuring that no group is disproportionately flagged as positive when the true outcome is negative.
- **Positive Class Balance:** Compares the distribution of predicted probabilities among individuals whose true outcome is positive between groups, ensuring that the model does not favor one group over another in its positive predictions.
- **Negative Class Balance:** Compares the distribution of predicted probabilities among individuals whose true outcome is negative between groups, ensuring that the model does not favor one group over another in its negative predictions.

Sufficiency

- **Predictive Parity:** Compares positive predictive values across groups, assessing whether the precision of positive predictions is equivalent.

Other Criteria

- **Brier Score Parity:** Compares the Brier score—the mean squared error of probabilistic predictions—is similar across groups, indicating comparable calibration.
- **Accuracy Parity:** Compares the overall accuracy of a predictive model is equivalent across different groups.
- **Treatment Equality:** Compares the ratio of false negatives to false positives across groups, ensuring the balance of missed detections versus false alarms is consistent.

Evaluating Fairness Criteria

The primary input to the {fairmetrics} package is a data frame or tibble which containing the model's predictions, true outcomes, and the protected attribute in question. Figure 1 shows the workflow for using {fairmetrics}. It is possible to evaluate a model for a specific or multiple group fairness metrics.

A simple example of how to use the {fairmetrics} package is shown below. The example makes use of the `mimic_preprocessed` dataset, a pre-processed version of the the

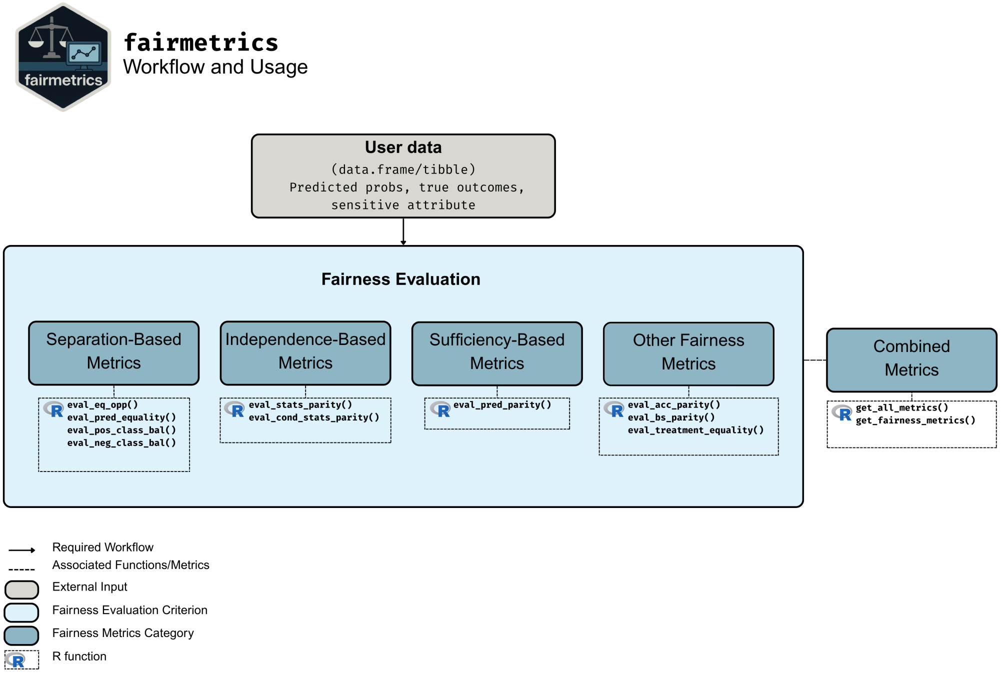

Figure 1: Workflow for using {fairmetrics} to evaluate model fairness across multiple criteria.

Indwelling Arterial Catheter (IAC) Clinical Dataset, from MIMIC-II clinical database[1] (J. Raffa 2016; J. D. Raffa et al. 2016). This dataset consists of 1776 hemodynamically stable patients with resperatory failure, and includes demographic information (patient age and gender), vital signs, laboratory results, whether an IAC was used, and a binary outcome indicating wheter the patient died within 28 days of admission.

While the choice of fairness metric used is context dependent, we show all metrics available with the `get_fairness_metrics()` function. In this example, we evaluate the model's fairness with respect to the protected attribute `gender`. For conditional statistical parity, we condition being `>=60` years old. The model is trained on a subset of the data, and predictions are made on a test set.

```r
library(fairmetrics)
library(dplyr)
library(magrittr)
library(randomForest)

# Load the example dataset
data("mimic_preprocessed")

# Split the data into training and test sets
train_data <- mimic_preprocessed %>%
  dplyr::filter(dplyr::row_number() <= 700)

test_data <- mimic_preprocessed %>%
  dplyr::mutate(gender = ifelse(gender_num == 1, "Male", "Female")) %>%
  dplyr::filter(dplyr::row_number() > 700)
```

[1] The raw version of this data is made available by PhysioNet (Goldberger et al. 2000) and can be accessed in {fairmetrics} package by loading the `mimic` dataset.

```
# Train a random forest model
rf_model <- randomForest::randomForest(
  factor(day_28_flg) ~ .,
  data = train_data,
  ntree = 1000
  )

# Make predictions on the test set
test_data$pred <- predict(rf_model, newdata = test_data, type = "prob")

# Evaluate predictive equality
# (Setting message=FALSE to avoid cluttering the output)

get_fairness_metrics(
 data = test_data,
 outcome = "day_28_flg",
 group = "gender",
 group2 = "age",
 condition = ">=60",
 probs = "pred",
 cutoff = 0.41
)

#>                                       Metric GroupFemale GroupMale Difference
#> 1                         Statistical Parity        0.17      0.08       0.09
#> 2  Conditional Statistical Parity (age >=60)        0.34      0.21       0.13
#> 3                          Equal Opportunity        0.38      0.62      -0.24 [
#> 4                        Predictive Equality        0.08      0.03       0.05
#> 5                 Balance for Positive Class        0.46      0.37       0.09
#> 6                 Balance for Negative Class        0.15      0.10       0.05
#> 7                          Predictive Parity        0.62      0.66      -0.04
#> 8                         Brier Score Parity        0.09      0.08       0.01
#> 9                    Overall Accuracy Parity        0.87      0.88      -0.01
#> 10                        Treatment Equality        1.03      3.24      -2.21 [
```

Among the fairness metrics calculated, metrics whose difference confidence intervals cross zero and whose ratio confidence intervals cross one indicate no significant difference between the groups. In this example, the statistical parity, conditional statistical pairty, equal opportunity, predictive equality, positive class balance, negative class balance, and treatment equaltiy metrics show a significant differences. While, the predictive parity, Brier score parity, and overall accuracy parity metrics do not show significant differences between the groups.

Should the user wish to calculate an individual metric, it is possible to use any of the `eval_*` functions. For example, to calculate the equal opportunity metric, the user can use the `eval_equal_opportunity()` function.

```
eval_eq_opp(
  data = test_data,
  outcome = "day_28_flg",
  group = "gender",
  probs = "pred",
  cutoff = 0.41
)
```

```
#>There is evidence that model does not satisfy equal opportunity.
#>  Metric GroupFemale GroupMale Difference    95% Diff CI Ratio 95% Ratio CI
#>1    FNR        0.38      0.62      -0.24 [-0.39, -0.09]  0.61 [0.44, 0.85]
```

For more traditional fairness metrics, such as the True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive Value (PPV), Negative Predictive Value (NPV), and others, the `get_all_metrics()` function can be used. This function returns a data frame with the metrics calculated for each group.

```
get_all_metrics(
  dat = test_data,
  outcome = "day_28_flg",
  group = "gender",
  probs = "pred",
  cutoff = 0.41
)

#>           Metric Group Female Group Male
#> 1            TPR         0.62       0.38
#> 2            FPR         0.08       0.03
#> 3            PPR         0.17       0.08
#> 4            PPV         0.62       0.66
#> 5            NPV         0.92       0.90
#> 6            ACC         0.87       0.88
#> 7    Brier Score         0.09       0.08
#> 8    FN/FP Ratio         1.03       3.24
#> 9 Avg Pred Prob         0.21       0.14
```

Related Work

Other R packages similar to {fairmetrics} include {fairness}(Kozodoi and V. Varga 2021), {fairmodels} (Wiśniewski and Biecek 2022) and {mlr3fairness}[mlr3fairness_package]. The differences between {fairmetrics} and these other packages is twofold. The primary difference between is that {fairmetrics} calculates the ratio and difference between group fairness criterion and allows estimated confidence intervals of fairness metrics via bootstrap - allowing for more meaningful inferences about the fairness metrics calculated. Additionally, in contrast to the {fairmodels}, {fairness} and {mlr3fairness} packages, the {fairmetrics} package does not posses any external dependencies and has a lower memory footprint, resulting in an environment agnostic tool that can be used with modest hardware and older systems. Table 1 shows the comparison of memory used and dependencies required when loading each library.

Package	Memory (MB)	Dependencies
fairmodels	17.02	29
fairness	117.61	141
fairmodels	58.11	45
fairmetrics	0.05	0

Table 1: Memory usage and dependencies of fairmetrics vs similar packages (MB)

For python users, the {fairlearn} library (Weerts et al. 2023) provides a broader set of fairness metrics and algorithms. The {fairmetrics} package is designed for seemless integration with R workflows, making it a more convenient choice for R-based ML applications.

Licensing and Availability

The {fairmetrics} package is under the MIT license. It is available on CRAN and can be installed by using `install.packages("fairmetrics")`. A more in-depth tutorial can be accessed at: https://jianhuig.github.io/fairmetrics/articles/fairmetrics.html. All code is open-source and hosted on GitHub. All bugs and inquiries can be reported at https://github.com/jianhuig/fairmetrics/issues/.

References

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, Massachusetts: The MIT Press.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research* 50 (1): 3–44. https://doi.org/10.1177/0049124118782533.

Castelnovo, Alessandro, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. "A Clarification of the Nuances in the Fairness Metrics Landscape." *Scientific Reports* 12 (1). https://doi.org/10.1038/s41598-022-07939-1.

Gao, Jianhui, Benson Chou, Zachary R. McCaw, Hilary Thurston, Paul Varghese, Chuan Hong, and Jessica Gronsbell. 2024. "What Is Fair? Defining Fairness in Machine Learning for Health." *arXiv.org*. https://arxiv.org/abs/2406.09307.

Goldberger, Ary L., Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals." *Circulation [Online]* 101 (23): e215–20. https://doi.org/10.1161/01.CIR.101.23.e215.

Grote, Thomas, and Geoff Keeling. 2022. "Enabling Fairness in Healthcare Through Machine Learning." *Ethics and Information Technology* 24 (3): 39. https://doi.org/10.1007/s10676-022-09658-7.

Kozodoi, Nikita, and Tibor V. Varga. 2021. *Fairness: Algorithmic Fairness Metrics*. https://CRAN.R-project.org/package=fairness.

Raffa, Jesse. 2016. "Clinical Data from the MIMIC-II Database for a Case Study on Indwelling Arterial Catheters (Version 1.0)." https://doi.org/10.13026/C2NC7F. https://doi.org/10.13026/C2NC7F.

Raffa, Jesse D., Mohammad Ghassemi, Tristan Naumann, Mengling Feng, and Daniel J. Hsu. 2016. "Data Analysis." In *Secondary Analysis of Electronic Health Records*, 109–22. Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_9.

Rajpurkar, Pranav, Emma Chen, Oishi Banerjee, and Eric J. Topol. 2022. "AI in Health and Medicine." *Nature Medicine* 28 (1): 31–38. https://doi.org/10.1038/s41591-021-01614-0.

Weerts, Hilde, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. "FairLearn: Assessing and Improving Fairness of AI Systems." *arXiv.org*. https://arxiv.org/abs/2303.16626.

Wiśniewski, Jakub, and Przemysław Biecek. 2022. "Fairmodels: A Flexible Tool for Bias Detection, Visualization, and Mitigation in Binary Classification Models." *The R Journal* 14 (1): 227–43. https://doi.org/10.32614/RJ-2022-019.